DOCUMENT RESUME

ED 389 515 SE 056 782

AUTHOR Bianchini, Julie; And Others
TITLE Cooperative Learning in the Untracked Middle School Science Classroom: A Study of Student Achievement.
PUB DATE 19 Apr 95
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 17-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement; Biology; *Cooperative Learning; Intermediate Grades; Junior High Schools; Middle Schools; Science Education
IDENTIFIERS *Middle School Students

ABSTRACT

In recent years, science educators have struggled with the dual challenge of equity and excellence, of providing all students access to a quality science education, and ensuring each student achieves an understanding of scientific concepts and processes. The Program for Complex Instruction, a model of group work, addresses this dual challenge by teaching various subject matter at a high intellectual level to students with diverse academic skills. By examining science learning in Complex Instruction classrooms, insight into the following three questions was sought: (1) What scientific facts, concepts, and applications do students learn during group work?; (2) Are all students provided equal access to science learning?; and (3) What are the strengths and limitations of using open-ended paper-and-pencil tests to measure student learning of science? The study was conducted in 13 middle school science classrooms over the course of 2 years. The results are organized into four areas of interest: conditions for learning, excellence, equity, and issues of test construction. Findings include: all observed classrooms met the standard for the percentage of students talking and working together; significant gains were found on all five pre-, post-tests concerning the topics of systems, respiration, digestion, circulation; reading scores were significantly correlated with the pre- and post-test scores on each of the five unit tests; gender effects were found on the various tests but they varied by grade level; and on some tests (respiration and circulation 2) students scored higher when tests lacked large percentages of drawings or diagrams while other tests (digestion and circulation 1) showed gains when large amounts of diagrams were included. Contains 19 references. (MVL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 515

COOPERATIVE LEARNING IN THE UNTRACKED MIDDLE SCHOOL SCIENCE CLASSROOM: A STUDY OF STUDENT ACHIEVEMENT

BY

JULIE BIANCHINI, NICOLE HOLTHUIS, AND KATHERINE NIELSEN
STANFORD UNIVERSITY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Julie A. Bianchini
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
APRIL 19, 1995

BEST COPY AVAILABLE

Cooperative Learning in the Untracked Middle School Science Classroom: A Study of Student Achievement

---

In recent years, science educators have struggled with the dual challenge of equity and excellence, of providing all students access to a quality science education, and ensuring each student achieves an understanding of scientific concepts and processes (American Association for the Advancement of Science, 1989; California Department of Education, 1990; National Research Council, 1994). Traditional science instruction often marginalizes those students without the expected repertoire of experiences, interests, and skills (Carter, 1990; Delpit, 1988; O'Loughlin, 1992). Moreover, statistics from national science exams (Educational Testing Service, 1992) as well as studies of student misconceptions of biological and physical phenomena (Bishop and Anderson, 1990; Bar, 1989; Driver and Easely, 1978) seem to confirm that few students achieve a coherent and useful understanding of science.

The Program for Complex Instruction, a model of groupwork developed by Elizabeth Cohen and colleagues at Stanford University, addresses this dual challenge by teaching various subject matter at a high intellectual level to students with diverse academic skills. Grounded in sociological theories of expectation states and organizational behavior, Complex Instruction is designed for use in academically, linguistically, and culturally diverse classrooms. The Program includes three components: group tasks requiring multiple intellectual abilities and centered around a big idea; a classroom management system of cooperative norms and procedural roles; and treatments used by the teacher to equalize rates of student interaction. The program attempts to achieve two goals: to foster higher-order, or conceptual thinking through open-ended, problem-solving group tasks; and to facilitate all students' access to materials, to group discourse, and thus, to learning.

Given the congruency in goals between science education and the Program for Complex Instruction, the purpose of our study is to examine science learning in Complex Instruction classrooms. Specifically, we constructed, administered, scored, and analyzed pre and post science unit tests. In doing so, we hope to provide insight into the following questions: (1) What scientific facts, concepts, and applications do students learn during groupwork? (2) Are all students provided equal access to science learning? (3) What are the strengths and limitations of using open-ended paper-and-pencil tests to measure student learning of science?

Theoretical Framework

Below, we provide an overview of the sociological framework of Complex Instruction. Such information serves to situate our work in a larger theoretical and research tradition. For our study, we have only drawn upon certain aspects of this framework.

Complex Instruction is grounded in a sociological framework of organizational theory and expectation states. It uses the principles of organizational theory to understand the relationships among student interaction, task structure, and achievement (Cohen, Lotan, and Leechor, 1989). When there are open-ended tasks in an organization, the greater the lateral communication among workers, the greater the productivity. Lateral communication helps workers cope with uncertainty (Perrow, 1967), increases the amount of information processed (Galbraith, 1973), and fosters higher-level search procedures for solving problems (March and Simon, 1958). Similarly, when there are open-ended group tasks in a classroom, the more students talk and work together, the more they learn (Cohen, 1984; Cohen, Lotan, and Leechor, 1989; Lotan, Cohen, and Holthuis, 1994).

Complex Instruction also draws from the principles of organizational theory to maximize student-student interaction, that element of groupwork most directly linked

to learning. In organizations, managers delegate authority to workers to enhance lateral communication and increase effectiveness. Similarly, in classrooms, teachers refrain from directly supervising student behavior and progress during groupwork. Instead, to increase student-student interaction, they delegate authority to the groups through the use of cooperative norms and procedural roles.

Complex Instruction applies the lens of status characteristics and expectation states theory to devise strategies to help equalize rates of interaction among students within a cooperative group. According to expectation states theory, an individual's access to materials, participation, and influence in a group is determined by his or her status (Berger, Cohen, and Zelditch, 1966). Examples of status characteristics that operate in classrooms include academic ability, gender, ethnicity, social class, and popularity. Students of high status expect and are expected to excel at the group task. They talk a great deal and their suggestions carry weight. In contrast, students of low status have limited access to group materials and discourse. Because they talk less than their high status counterparts, they learn less (Cohen, 1984).

Objectives

To explore student learning of science in Complex Instruction classrooms, we used classroom observations and open-ended, content referenced unit tests. First, we statistically analyzed classroom observations to answer the following question: Did classrooms meet Complex Instruction's implementation standards? Empirical research by colleagues at Complex Instruction has shown a relationship between the quality of groupwork and student achievement (Cohen, Lotan, and Leechor, 1989). In a classroom, if the average student on-task talk is low and/or disengagement high, performance on achievement tests suffers. Based on this relationship, researchers have identified implementation standards for student behavior: classrooms that do not meet these standards have significantly lower achievement gains. Thus, we expect students

in classrooms that meet these implementation standards–classrooms with high on-task talk and low disengagement–to perform well on these content-referenced tests.

Second, what kinds of scientific concepts and skills did students learn? Given the theoretical link between open-ended tasks and student-student communication, the Program for Complex Instruction creates group activities that are open-ended and designed to stimulate higher-order thinking. We thus expect students to excel on test questions that require application, analysis, and/or synthesis of concepts and ideas.

Third, did student learning of scientific concepts and processes differ along lines of gender or conventional academic achievement? Starting from expectation states theory and expanding into the larger arena of access, Complex Instruction attempts to provide students access to learning through a curriculum requiring multiple intellectual abilities and through pedagogical techniques designed to equalize rates of student interaction. In this study, we examine two of many factors that restrict students' access to learning: previous academic achievement (indicated by national reading percentile scores) and gender. We expect to see little difference in achievement gains between students who do well and students who do poorly in traditional academic settings. We also expect to see no difference in achievement gains between boys and girls.

Fourth, what did we learn about test construction and scoring? Tests developed for this study included both open-ended and multiple choices questions as well as diagrams and illustrations. We expect to learn how to eliminate gratuitous difficulties, refine test questions, and better tap students' understandings through examination of student responses and reflection on our own scoring process.

Sample, Setting, and Methods

We conducted our study in thirteen middle school science classrooms over the course of two years. During the 1992-1993 academic year, approximately 260 sixth and eighth grade students in ten classrooms participated, and during 1993-1994,

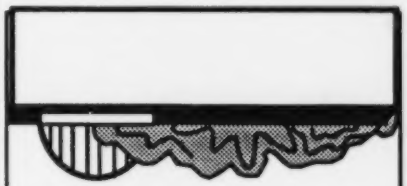
approximately 80 sixth grade students in three classrooms. These students were taught by four science teachers, each experienced in teaching science and trained in Complex Instruction. They attended two middle schools in the greater San Francisco Bay Area. Both schools had an academically, linguistically, and ethnically diverse student population. Both had made a commitment to untrack in all subjects.

To ensure classrooms met Complex Instruction's standards of implementation, researchers conducted systematic observations using an instrument we call the Whole Class Instrument in seven of the thirteen classes in our study. These researchers were members of the Complex Instruction staff, had achieved a reliability of 90% with the Whole Class criterion scorer during the 1992-1993 academic year, and had used the instrument in other Complex Instruction research endeavors. Researchers completed a minimum of 10 observations per classroom while students were engaged in groupwork. Data collected with this instrument provided snapshots of student behavior in groups. Subsequently, the data were used to determine an average percentage of students talking and working together, and an average percentage of students disengaged.

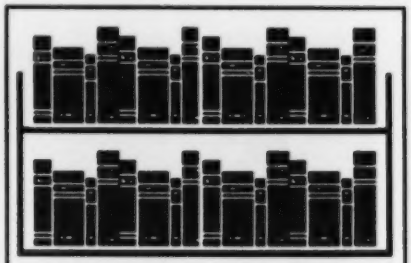
To assess students' factual knowledge, conceptual understanding, and ability to apply and synthesize scientific information, we developed tests for each of four Complex Instruction units: Circulation, Respiration, Digestion, and Systems. Tests contained six to ten questions with many subparts. The questions ranged in format from multiple choice to short answer to the drawing of a diagram. They ranged in level from recall to synthesis. Many included illustrations to facilitate student comprehension. Below, in Figures 1 and 2, we provide examples of test questions.

6. Your younger sister asks, "What does a digestive system do?" Use an analogy to help explain the purpose of the digestive system to your sister.

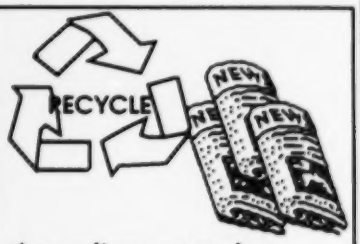
a. Circle **the one best** analogy from those given below. The digestive system is like . . .



A street sewer into which rain water drains and flows out to a river.

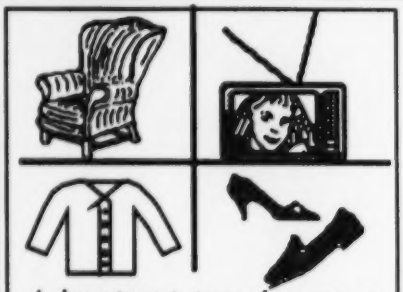


A library from which books are borrowed and returned.



A recycling center where newspapers are collected, shredded and reused.

A bakery where bread and cookies are made and sold.



A department store where many different items are found.

b. Explain why you chose this analogy.

Figure 1. *Question 6 in DigestionTest.*

5. Match the following parts of the circulatory system to their function. There are more choices than needed.

___	Carries oxygen and nutrients to cells, carbon dioxide and waste from cells.	a.	Heart
___	A small vessel that allows the exchange of substances between blood and cells.	b.	Blood
___	A large vessel that transports blood from body to heart.	c.	Vein
___	Pumps blood throughout the body.	d.	White blood cell
___	A large vessel that takes blood from heart to organs.	e.	Arteriole
___	Contains proteins that transport oxygen.	f.	Capillary
		g.	Red blood cell
		h.	Artery

Figure 2. *Question 5 in Circulation 2 test.*

Tests were administered to classrooms on a unit-by-unit basis: a pretest was given prior to implementation, and a posttest, upon unit completion. During the 1992-1993 school year, sixth and eighth grade students completed one to three Complex Instruction tests: Systems, Circulation 1, and/or Digestion. The following school year, sixth grade students completed two Complex Instruction unit tests: Circulation 2 (a revised Circulation 1 test) and Respiration .

Researchers scored these unit tests in two-to-three hour blocks over the course of one year. For a given unit test, we scored the closed-ended questions (multiple choice, matching, fill-in-the-blank) first. We then moved to the scoring of open-ended questions, that is, questions that required students to respond in their own words. For each open-ended question, we constructed a scoring rubric and obtained 80% or greater inter-scorer reliability.

Once all scoring was completed, researchers proceeded with analyses of test results on both quantitative and qualitative fronts. Percentage pre, post, and gain scores were statistically analyzed. In addition, student responses were studied qualitatively for evidence of scientific understanding and for flaws in test construction.

Finally, to ascertain differences in student performance by gender and previous academic achievement, researchers collected two additional pieces of information: students' gender and most recent national reading percentile score from a standardized test. (Most standardized test scores were from the CTBS; a few, from MAT-6.) We used this information to analyze test performance by gender and previous reading achievement.

Results

The following results are organized around our four areas of interest: the conditions for learning, excellence, equity, and issues of test construction.

Conditions for Learning

To check that all classrooms met the minimum Complex Instruction implementation standards, we analyzed the Whole Class observational data by classroom. In all, observations were made in seven of the thirteen classes tested. Averages were calculated for the percentage of students talking and working together and the percentage disengaged in each class. (See Table 1 below.)

Table 1
Complex Instruction implementation standards by class

	Ave. % of Students Talking and Working	Average % of Students Disengaged
Class 1	35.0	22.0
Class 2	47.0	8.0
Class 3	41.0	4.4
Class 4	37.2	4.8
Class 5	43.4	9.9
Class 6	54.6	1.6
Class 7	37.1	9.8

Based on these percentages, all of the classes met the implementation standard for the percentage of students talking and working together ($\geq$35%). All but Class 1 satisfied the implementation standard regarding the percentage of students disengaged ($\leq$14%). As a result, the quality of groupwork implementation in these seven classes was considered to be relatively good. Subsequent analyses were completed using the total data sample which includes these seven classes as well as six others in which tests were administered but no observational data was collected.

Excellence: What Did Students Learn?

Analyses were performed to determine the average pre, post, and gain scores on each of the five tests.

The five tests were not of equal value (a total of 56 points was possible on the Systems, 43 on the Respiration, 51 on the Digestion, 93 on Circulation 1, and 83 on the Circulation 2). Thus, in order to compare scores across tests, percentage totals were calculated for the pre, post, and gain scores by dividing the test scores by the total number of points possible. As a result, the percentage scores were simply a linear transformation of the raw scores–analyses such as correlations and *t*-tests were not affected by this recalculation. Table 2 provides a summary of the percentage pretest, posttest, and gain scores for each test.

Table 2
Descriptive statistics for percentage pre, post, and gain scores on each test

	Systems Test			Respiration Test			Digestion Test		
	% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain
N	206	206	206	65	65	65	172	172	172
Mean	36.0	60.9	24.9	27.4	42.2	14.8	35.4	42.7	7.3
SD	16.2	18.9	17.4	13.3	19.1	11.8	16.2	16.8	10.8
Min	1.8	7.1	-28.6	3.5	7.0	-8.1	3.9	0.0	-21.6
Max	76.8	96.4	85.7	77.9	84.9	44.2	77.5	85.3	34.3

	Circulation Test 1			Circulation Test 2		
	% Pre	% Post	% Gain	% Pre	% Post	% Gain
N	135	135	135	69	69	69
Mean	21.6	35.6	14.0	17.3	36.1	18.8
SD	10.2	14.4	10.5	8.2	18.1	12.8
Min	4.8	7.8	-7.8	2.4	8.4	-9.6
Max	55.4	82.5	45.8	42.2	83.7	53.6

T-tests indicated that the posttest scores were significantly higher than the pretest scores for each of the five tests ($p<0.001$ for each). The average scores on the Systems pretest (36.0%) and posttest (60.9%) were the highest of the five tests. Learning gains were greatest for the Systems test (24.9%) and lowest for the Digestion test (7.3%).

Next, we asked whether students learned what we wanted them to learn. That is, did students show significant learning gains on higher-order concepts and processes? Analysis of students' scores on higher-order questions provided us with some answers.

We categorized higher-order questions as those questions that asked students to apply, analyze, and/or synthesize scientific knowledge. For each test, we aggregated student scores on these questions and then computed percentage pre, post, and gain scores by dividing each student's total score on all higher-order questions by the total number of points possible on these type of questions (24 points were possible on the Systems test, 17 on Respiration, 27 on Digestion, 38 on Circulation 1, and 34 on Circulation 2).

Table 3 summarizes the percentage pre, post, and gain scores on higher-order questions. In addition, we have noted the *t* values of the dependent *t*-test run on the pre and posttest scores. Shaded cells indicate a statistically significant difference between the pre and post test scores.

Table 3
Descriptive statistics for percentage pre, post, and gain scores on higher-order questions

		Systems Test	Respiration Test	Digestion Test	Circulation Test 1	Circulation Test 2
Pre	X	**37.8**	**46.5**	**34.3**	**25.1**	**23.4**
	(SD)	**(23.7)**	**(21.0)**	**(20.2)**	**(15.0)**	**(12.2)**
Post	X	**59.6**	**64.7**	**39.7**	**38.3**	**38.1**
	(SD)	**(23.9)**	**(27.9)**	**(21.2)**	**(17.9)**	**(19.6)**
Gains	X	**21.8**	**18.1**	**5.6**	**16.9**	**14.7**
	(SD)	**(23.3)**	**(23.0)**	**(15.0)**	**(13.8)**	**(14.0)**
t value		-13.439***	-6.383***	-4.875***	-10.201***	-8.777***

*** $p<0.001$

Trends in higher-order scores mirror those identified for the overall scores. The posttest scores on the higher-order questions were significantly higher than the pretest scores for each of the five tests as determined by the *t*-tests. Gains were largest on the Systems test (students scored 21.8% better on the posttest). Students showed the smallest gain in higher-order thinking on the Digestion test (5.6%).

In addition to our quantitative analyses, we conducted qualitative analyses of student responses. We found trends in the following three areas: (1) misconceptions about scientific content and processes; (2) ability or inability to draw scientific diagrams; and, (3) ability or inability to construct reasoned arguments.

Students' responses revealed four misconceptions. One, a small number of students confused the consequences of a phenomena with its cause. For example, in the Circulation test, students were asked to explain two possible consequences of a heart attack. Students gave answers such as, "If you have high blood pressure and if you don't eat healthy food your heart will clog up."

Another misconception was also found in the Circulation tests. Many students thought it was important that blood types be the same in a transfusion because of the possibility of contracting AIDS or another disease. In one case, a student wrote, "If he loses a lot of blood he can die and blood transfusion can give him HIV."

In the Systems test, a third misconception involved the scientific idea of a collection, as opposed to a system. Many students deemed a representation of a system a collection because they saw it as something that is collected. For example, one student said a flower growing in a pot was a collection, "Because people collect plants." Another student called a dinosaur playing a guitar a collection, "Because there are a bunch of kids that collect dinosaurs."

In contrast to the above examples of students' misconceptions of content, a final misconception revolved around scientific process: the majority of students did not understand or use the idea of an experimental control. On two different exams, students were asked devise an experiment. Their answers included the following confusions about a control: "Our control is our brain," "My control is to get the breathing rate," and "Go into the school and have a sign-up on who would want to be like the people in the TV show" (The first two answers are in response to a question on breathing rate; the last answer, in response to a question about the effect of a TV show

on adolescent eating habits and body image.) Eighth graders' understanding of a control was slightly more sophisticated. Indeed, one eighth grader wrote: "Have one person watch *Rock High* for one year. Have another person not watch *Rock High* for one year. At the end of one year see how they both turn out." Another suggested scientists, "Survey high school kids that watch and don't watch the show."

Second, review of student responses also revealed that many students had difficulty producing a scientific diagram. On the Circulation tests, students were asked to explain the path that oxygen takes through the circulatory system, using a diagram to help. Almost all students were unable to provide the correct path. However, they did seem to have some idea of what constitutes a diagram, e.g. drawings, labels, and/or arrows. In each of the six classes tested, between a sixth and a third of the students provided such diagrams.

On the Respiration test, a question required two diagrams and specifically asked the students to include labels. In this case, approximately half of the students provided diagrams. Below their diagrams, students were asked to provide similar information, but in written form. Students generally attempted to answer both the diagram and written part. In addition, there was a significant correlation ($p<0.001$) between these two parts of the question.

Third, students had mixed success in providing reasoned arguments in support of their answers. Across three questions, a trend in student abilities emerged: students had the most difficulty supporting their personal opinion, the least difficulty supporting their answer through recall of information, and supporting a scientific analogy fell somewhere in between. For example, in the Circulation tests, one question asked, "Do you think ads, such as the one shown above [an ad advertising smoking and drinking at a baseball game], affect people's attitudes or behaviors about health? Explain." Students found it difficult to support their personal opinion; the majority of students basically ignored the thrust of the question and instead focused on drinking and smoking as unhealthy behaviors. From

Digestion, a question called upon students to explain why the foods they had chosen made up a healthy lunch. Most commonly, students responded that their lunch met all of the food groups, was low in fat, and provided protein and vitamins. And a third question, also in Digestion, presented students with five different illustrated analogies; students were to pick the one that best fit the digestive system and logically support their choice. The question yielded a wide range of answers. There were several strong answers such as "I chose the recycling one because it's the best one. You eat food, it gets digested, and parts are reused." However, many students did not fully support their choice.

Equity: Who Learned?

The test scores were analyzed to see if students' reading ability and gender affected their performance.

Effect of Reading Ability

National curve equivalents (NCEs) were computed using the national percentile reading scores collected for each student. These NCEs were correlated with test pre, post, and gain scores. (See Table 4 below.)

Table 4.
Correlation coefficients: NCE Scores and Percentage Pre, Post, and Gain Scores

	Systems			Respiration			Digestion			Circulation 1			Circulation 2		
	% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain
NCE Scores	0.479 ***	0.503 ***	0.079	0.550 ***	0.710 ***	0.474 ***	0.698 ***	0.692 ***	0.027	0.622 ***	0.619 ***	0.265 **	0.774 ***	0.684 ***	0.467 ***

** $p<0.01$
*** $p<0.001$

Reading scores were significantly correlated with the pre and post test scores on each of the five unit tests. Reading scores were also significantly correlated with percentage gain scores for three of the five tests–Respiration, Circulation 1, and Circulation 2.

We completed additional analysis to examine the effect of students' reading ability on their achievement by organizing students into two groups based on their national reading percentile score. The test scores of students reading at or above grade level (reading scores at or above the 40th percentile) were compared to those of students reading below grade level (reading scores below the 40th percentile). Table 5 indicates the average percentage pre, post, and gain scores by each group. We have also included the *t* values for independent *t*-tests of scores for students reading below grade level versus those reading at or above grade level. Shaded cells indicate a statistically significant difference in the scores of these two groups.

Table 5
Percentage pre, post, and gain scores for students reading below grade level and those reading at or above grade level

		Systems Test			Respiration Test			Digestion Test		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain
Low Reading	X (SD)	**29.7 (12.3)**	**54.1 (16.0)**	24.3 (16.5)	**23.7 (9.3)**	**34.7 (14.8)**	**11.0 (10.0)**	**24.3 (11.3)**	**31.6 (13.6)**	7.3 (10.7)
High Reading	X (SD)	**41.7 (16.6)**	**68.4 (16.7)**	26.7 (18.7)	**34.7 (14.4)**	**55.4 (14.9)**	**20.7 (12.0)**	**44.1 (14.8)**	**51.2 (14.6)**	7.0 (10.4)
t value		5.360***	5.749***	0.880	3.486***	5.197***	3.253**	3.926***	3.338***	-0.149

		Circulation Test 1			Circulation Test 2		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain
Low Reading	X (SD)	**16.5 (7.6)**	**28.8 (10.4)**	**12.3 (9.0)**	**13.3 (5.3)**	**28.4 (13.4)**	**15.1 (10.5)**
High Reading	X (SD)	**26.5 (8.8)**	**43.4 (13.7)**	**16.9 (11.4)**	**22.7 (7.3)**	**47.3 (17.0)**	**24.7 (14.0)**
t value		6.504***	6.416***	2.376**	5.802***	4.848***	3.018**

** $p<0.01$
*** $p<0.001$

T-tests showed that students with higher reading ability did significantly better on all the tests, both pre and post ($p<0.001$). Students of high reading ability had higher gain scores than students of low reading ability on both the Circulation and Respiration unit tests. Gain scores between the two groups did not vary significantly on either the Systems or Digestion tests.

Effect of Gender

Test scores were analyzed to determine the effect of gender on student scores. Did girls and boys do equally well on these tests? Does a gender gap exist in 6th grade? Does it begin to appear in the 8th grade (recall, two tests–Systems and Digestion–were administered in both the 6th and 8th grades)? The table below, Table 6, provides the percentage pre, post, and gain scores of boys and girls for each test. Again, the shaded cells indicate a statistically significant difference between the scores of boys and girls.

Table 6
Percentage pre, post, and gain scores for boys and girls (total sample, 6th grade, and 8th grade)

			Systems Test			Respiration Test			Digestion Test		
			% Pre	% Post	% Gain	% Pre	% Post	% Gain	% Pre	% Post	% Gain
Total	Girls	X (SD)	**38.1 (16.5)**	59.7 (19.9)	**21.6 (17.0)**	not applicable (test was administered to 6th grade students only)			**38.0 (16.6)**	**46.1 (15.7)**	8.1 (9.8)
	Boys	X (SD)	**33.1 (15.1)**	60.6 (18.2)	**27.5 (17.2)**				**32.0 (16.0)**	**39.2 (17,5)**	7.2 (11.9)
	t value		-2.152*	0.317	2.349*				-2.137*	-2.428**	-0.509
6th	Girls	X (SD)	35.5 (15.5)	57.9 (19.9)	22.4 (17.4)	25.8 (10.6)	42.5 (17.0)	16.6 (11.4)	**30.8 (12.9)**	**39.7 (12.2)**	8.8 (11.3)
	Boys	X (SD)	32.4 (14.2)	57.8 (16.7)	25.4 (16.7)	29.0 (15.6)	41.9 (21.4)	12.9 (12.1)	**22.7 (10.4)**	**28.9 (13.1)**	6.2 (11.3)
	t value		-1.171	-0.014	1.006	-0.969	0.118	1.293	-3.116**	-3.858***	-1.095
8th	Girls	X (SD)	**42.9 (17.3)**	63.2 (19.8)	**20.3 (16.3)**	not applicable			**51.1 (14.6)**	58.0 (14.6)	6.9 (8.1)
	Boys	X (SD)	**34.9 (17.2)**	67.4 (20.1)	**32.6 (17.6)**				**43.9 (13.9)**	52.3 (13.0)	8.5 (12.7)
	t value		-1.792*	0.830	2.815**				-1.871*	-1.511	0.538

Table 6 (cont.)

			Circulation Test 1			Circulation Test 2		
			% Pre	% Post	% Gain	% Pre	% Post	% Gain
Total	Girls	X (SD)	not applicable			not applicable		
	Boys	X (SD)						
	t value							
6th	Girls	X (SD)	24.4 (10.0)	38.8 (14.6)	14.4 (10.5)	17.7 (7.4)	35.2 (16.1)	17.5 (11.6)
	Boys	X (SD)	18.0 (8.7)	31.5 (13.1)	13.5 (11.2)	17.0 (9.0)	37.0 (19.9)	20.0 (13.8)
	t value		-3.536**	-2.707**	-0.410	0.354	-0.423	-0.827
8th	Girls	X (SD)	not applicable			not applicable		
	Boys	X (SD)						
	t value							

* $p<0.05$
** $p<0.01$
*** $p<0.001$

T-tests were run in order to determine if the scores of girls and boys differed significantly. In 6th grade, girls scored significantly higher than boys on some of the pre and post tests (Circulation and Digestion tests), although gain scores for boys and girls did not differ significantly. However, in the 8th grade, boys gained significantly more than the girls ($p < .01$) on the Systems test: the boys had an average percentage gain of 32.6 ($SD = 17.6$), and the girls, of 20.4 ($SD = 16.3$).

Issues of Test Construction: What Did We Learn?

Analysis of data provided us with information not only about student performance but about construction and scoring of the test itself.

Type of Question: Scores on Questions with and without Illustrations

In designing the tests, we attempted to make them less reading dependent both by including diagrams and pictures in many of the test questions and by requiring

students to answer using a drawing or diagram. Did students score better on questions that included an illustration than those that did not?

The table below, Table 7, provides the percentage pre, post, and gain scores for the pictorial and non-pictorial questions on each test. The Systems test is not included in this analysis because all the items contained pictures or diagrams.

Table 7
Percentage pre, post, and gain scores on pictorial vs non-pictorial questions

		Respiration Test			Digestion Test		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain
Pictorial	X (SD)	**20.1 (11.4)**	**32.9 (16.4)**	12.8 (11.8)	**38.7 (18.1)**	**45.5 (18.4)**	6.8 (13.6)
Non-Pictorial	X (SD)	**33.6 (16.8)**	**49.2 (22.5)**	16.6 (15.6)	**25.4 (17.4)**	**34.4 (18.3)**	9.0 (16.1)
t values		8.402***	8.258***	1.412	-9.927***	-9.631***	1.362

		Circulation Test 1			Circulation Test 2		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain
Pictorial	X (SD)	**24.7 (12.6)**	**44.6 (17.5)**	**19.9 (15.8)**	**13.0 (6.4)**	32.0 (17.1)	**19.0 (14.5)**
Non-Pictorial	X (SD)	**19.8 (11.6)**	**30.2 (15.5)**	**10.4 (11.7)**	**18.2 (10.5)**	32.5 (17.9)	**14.3 (13.0)**
t values		-4.323***	-10.710***	-6.460***	4.689***	0.332	-2.505**

** $p<0.01$
*** $p<0.001$

On the pre and post tests for Respiration, students scored significantly higher on questions without illustrations. Just the opposite is true of the pre and post Digestion tests: students scored significantly higher on those questions with pictures or diagrams. Similarly, on Circulation 1 pre and post tests, students scored higher on the pictorial questions. They showed higher gains on those questions on the Circulation test 1 as well. On Circulation test 2, students scored higher on the non-pictorial questions on the pretest, though gains were higher on the pictorial questions.

Qualitative analysis also yielded ambiguous results. In some instances, illustrations seemed to constrain or confuse student responses. Many students

interpreted diagrams too literally. For example, we expected a drawing of a dinosaur playing a guitar to be interpreted as a system, yet several students gave answers similar to the following: "It is a collection because a rocking dinosaur could only be a toy or a doll so it could be a collectible toy" and "This is a collection. The dinosaur is dead and cannot play a guitar." Many students were also misled by illustrations that did not closely fit with the question. For example, one question asked students to describe the path of oxygen from the air to the cell but the accompanying diagram did not directly correspond: it was at the cellular level and did not include the air and the whole body. Thus, students received mixed messages about the focus of the question and may have mistakenly limited their answers.

In other instances, students clearly benefited from the acceptance of drawings as answers: they were better able to convey what they knew through an illustration than through words. For example, one question in the Circulation test asked students how they would measure certain vital signs. Some students scored high on this question by drawing stethoscopes, thermometers, and other medical tools. These same students scored very close to zero on the rest of the exam. It seems likely that these students were very poor writers, but this question enabled them to present some of their knowledge in pictorial form.

Finally, researchers found evaluating students' pictures incredibly difficult and time-consuming. With one question in particular, inter-scorer reliability was never achieved: two scorers scored all the exams together. Scoring the pictures was also frustrating: a sloppy picture that met our criteria could get the same or higher score than a carefully crafted one. Can and should scoring criteria include aesthetic as well as content standards? We reward traditional reading and writing abilities, why not artistic?

Structure of Question: Scores on Open- and Closed-Ended Questions

Scores were also analyzed by the *structure* of the question: open- or closed-ended. Open-ended questions were defined as those questions that asked students to draw or write their own answer. Closed-ended questions were defined as those that provided a list of possible answers from which students could choose. These questions included multiple-choice, fill-in-the-blank, and matching items.

Analysis of student scores on open- and closed-ended questions are reported in Table 8 below. The Systems test is not included as it did not contain any closed-ended questions.

Table 8
Percentage pre, post, and gain scores on open- and closed-ended questions

		Respiration Test			Digestion Test		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain
Open-Ended	X (SD)	**31.5 (14.5)**	**44.8 (19.6)**	**13.4 (13.5)**	**30.1 (18.5)**	**36.8 (20.1)**	6.7 (14.7)
Closed-Ended	X (SD)	**20.6 (15.7)**	**37.7 (21.4)**	**17.2 (15.1)**	**40.0 (17.3)**	**47.9 (17.0)**	7.9 (14.0)
t values		-6.214***	-4.071***	1.910*	8.681***	9.412***	0.791

		Circulation Test 1			Circulation Test 2		
		% Pre	% Post	% Gain	% Pre	% Post	% Gain
Open-Ended	X (SD)	**19.8 (10.7)**	35.7 (16.5)	**16.0 (12.6)**	**14.0 (8.4)**	**34.2 (20.0)**	**20.2 (14.8)**
Closed-Ended	X (SD)	**25.7 (12.7)**	35.3 (13.8)	**9.6 (13.4)**	**24.5 (11.7)**	**39.1 (15.6)**	**14.6 (13.7)**
t values		6.364***	-0.371	-4.665***	8.021***	3.133**	-3.127**

* $p<0.05$
** $p<0.01$
*** $p<0.001$

While students scored significantly higher on open-ended questions on the Respiration pre and post tests, they scored significantly higher on the closed-ended questions on the Digestion and Circulation 2 pre and post tests as well as the

Circulation 1 pretest. Posttest scores on the Circulation test 1, and gain scores on the Digestion test did not vary significantly by question type.

As with the illustrated questions, researchers found scoring open-ended questions time-consuming. In addition, scorers often strongly disagreed about when to take a student's response literally and when to read into it. For example, how should one interpret the following answer to a question asking about the importance of blood typing in transfusions: "Because his blood is one type and if you put another type in, his blood will go crazy." If such an answer were interpreted at its "maximum," that is with a belief that the student knew the correct response but was unable to or did not fully convey her knowledge, it would receive full credit. On the other hand, a strict interpretation would give it a lower score (which it received).

Questions that asked students to apply scientific information to real-life situations were often the most difficult to score. For example, scorers disagreed over what constituted the healthiest of lunches. Is Coke unhealthy? What about pizza? For young teens, is non-fat milk healthier than whole milk? Are some foods neutral–neither healthy or unhealthy? What if someone has a restricted diet and is coming from that vantage point? How does one include the important concept of moderation? Scorers reached an uneasy consensus on such important questions.

Implications

What did students learn? Students registered the highest learning gains–both overall and on higher-order questions–for the Systems test. Students also made significant gains on the other four tests, although their gains were noticeably smaller. The striking difference in student performance between the Systems test and the other four tests highlights the need for closer alignment between unit purpose, content taught, and test construction. The Systems unit was compact. Its purpose was to allow students to explore a few key concepts around the theme of systems over a small

number of group activities. In contrast, the other units were much larger in size and scope: there were both more concepts to cover and more ways to investigate them, e.g. text, experiments, and worksheets as well as group activities. In the Systems unit, there was also a tight match between concepts explored in the group activities and questions asked on the unit test. Again, in contrast, the other tests attempted to measure student learning of concepts and applications that were often addressed superficially or not at all over the course of the unit.

Did students learn *enough*? One way to ask this question is by considering student posttest scores: By the end of a given unit, did students obtain a reasonable understanding of the content taught? We cannot answer this question empirically–the data, in and of themselves, do not provide standards of worth. Unfortunately, the goals of the units were not sufficiently articulated to allow us, as researchers, to determine what is a reasonable understanding. A second way to phrase this question is by examining student gain scores: Were differences between pre and post tests satisfactory? Unfortunately, our study did not include a control group. (This was not by choice: no comparable classrooms were available for study.) We do not know if students would have learned more had they been taught a different way or with a different curriculum. Thus, while we struggled with these questions, they remain unanswered.

How could student learning be improved? Qualitative analyses of student responses offer three possibilities. First, we need to provide students explicit instruction and practice in designing experiments, drawing scientific diagrams, and constructing reasoned arguments. As Lemke (1990) argues, students do not come to the classroom with such skills–they must be initiated into the culture of science. Second, we need to better address gaps or confusions in students' scientific understanding during either whole class discussions or small group activities. Finally, from the beginning of a given unit, students need to be made aware of the purpose of the unit, the specific

content and processes they are expected to learn, and the standards by which they will be evaluated. Such explicit guidelines should facilitate students' learning by giving them goals to achieve and a means to focus their work.

Did all students have equal access to learning? As stated in the introduction, one goal of Complex Instruction is to provide all students equal access to learning, regardless of previous academic achievement or gender. Our goal in designing the unit tests, then, was to make them independent of reading, to include illustrations and open-ended questions, so that we could determine whether the curriculum and instructional strategy provided all students access to learning. Despite our efforts, there was a strong relationship between national reading percentile scores and pre and post test scores. There are three possible explanations for this disturbing finding: (1) poor readers could not read the test questions, (2) they could not express, in writing, what they knew, and/or (3) they could read and write sufficiently for the test but lacked the necessary "cultural capital" to perform well , e.g., they are not savvy in test-taking or have had little out-of-school exposure to science.

The fact that reading ability is correlated with test performance makes it difficult to interpret differences in test results between low and high readers. Students with low reading ability not only scored significantly lower than the high readers on the pre and post tests, but also gained significar less on three of these tests. One explanation is that the differences in gains are artifacts of the tests: low readers registered lower gain scores because they could not read or respond to many of the questions. Another explanation is that such differences are due to real differences in learning: the rich got richer, the poor, poorer. If so, lower gain scores by low readers raise concerns about the curriculum and instructional strategy. Either the group activities did not meet Complex Instruction's criteria of open-endedness and multiple intellectual abilities, or these qualities are not sufficient to promote equal access to scientific discourse and understanding.

To determine if Complex Instruction curriculum materials and instructional strategy provide equal access to science learning, future studies should attempt to better understand the effects of reading ability on test performance. We need to use multiple forms of assessment, e.g., portfolios, interviews, and/or performance-based tests as well as paper-and-pencil tests. Of course, this suggestion is not necessarily practical.

With regard to issues of gender equity, our results are more encouraging. Girls did not do worse than boys on any of the pre or post tests. Indeed, in some cases, they did significantly better. However, boys gained significantly more than girls on some tests. All in all, there are no clear trends to interpret.

What did we learn about test construction and scoring? We included both pictorial and open-ended questions in hopes of facilitating students' understanding of questions and their ability to respond to them. Unfortunately, this strategy did not appear to work: students did not perform consistently better on either type of question. At times, the inclusion of pictorial and open-ended questions aided students and, at other times, hindered them. One conclusion is that we need to pay careful attention to the kinds of illustrations provided in an exam and the ways they may mislead students. A second conclusion is that students did not find interpreting and drawing illustrations as easy as we thought they would. As with all skills, they need explicit instruction and practice.

We do not recommend, however, the elimination of either pictorial or open-ended questions. For example, although students did not perform consistently better on open-ended questions and scoring these questions was time consuming, such questions did provide numerous insights into ways to improve both the curriculum and the tests. We would have gained little information about student misconceptions if the tests had only consisted of multiple choice or fill-in-the-blank questions. We also would have failed to debate conceptions of what is healthy or to consider the importance of rewarding other abilities besides reading and writing. Moreover, we would not have

understood the ways illustrations can both facilitate and constrain a student's ability to respond to a given question.

Ultimately, our findings do not suggest all that we had hoped. However, they do provide insight into how curriculum and instruction can be better tailored to meet the goals of equity and excellence in science education. They also underscore the difficulties in separating the effects of a curriculum and instructional approach from the limitations of paper-and-pencil assessment. Lastly, our findings highlight the need to explicitly teach students about science, how to do science, and how to convey said knowledge.

References

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, D. C.

Bar, V. (1989). Children's views about the water cycle. *Science Education, 73,* 481-500.

Berger, J. B., Cohen, B. P., and Zelditch, M. Jr. (1966). Status characteristics and expectation states. In Joseph Berger and Morris Zelditch, Jr. (Eds.), *Sociological theories in progress.* (Vol. 1, pp. 29-46). Boston: Houghton-Mifflin.

Bishop, B., and Anderson, C. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27,* 415-427.

California Department of Education. (1990). *Science Framework.* Sacramento, CA.

Carter, Carolyn. (1990). Gender and equity issues in science classrooms: Values and curricular discourse. In Don Emil Herget's (Ed.), *Papers from Proceedings of the First Annual Conference of the History and Philosophy of Science in Science Teaching* (pp. 122-132). University of Florida, Tallahassee, Florida.

Cohen, E. G. (1984). Talking and working together: Status, interaction, and learning. In P. Peterson, Louise C. Wilkinson, and Maureen Hallinin (Eds.), *Instructional groups in the classroom: Organization and processes* (pp. 171-188). New York: Academic Press.

Cohen, E. G., Lotan, R. A., and Leechor, C. (1989). Can classrooms learn? *Sociology of Education, 62,* 75-94.

Delpit, Lisa. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review, 58,* 280-298.

Driver, R. and Easely, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education, 5,* 61-84.

Educational Testing Service. (1992). *The 1990 Science Report Card: NAEP's assessment of fourth, eighth, and twelfth graders.* Washington, D. C.: Office of Educational Research and Improvement, U.S. Department of Education.

Galbraith, J. (1973). *Designing complex organizations.* Reading, MA: Addison-Wesley Publishing Co.

Lemke, J. (1990). *Talking science: Language, learning, and values.* Norwood, New Jersey: Ablex Publishing Company.

Lotan, R. A., Cohen, E. G., and Holthuis, N. C. (1994). Talking and working together: Conditions for learning in Complex Instruction. Paper presented at Annual Meeting of American Educational Research Association. New Orleans, LA.

March, J. G., and Simon, H. A. (1958). *Organizations*. New York: John Wiley & Sons.

National Research Council. (1994, November). *National Science Education Standards: Draft*. Washington, D. C.: National Academy Press.

O'Loughlin, Michael. (1992). Rethinking science education: Beyond Piagetian constructivism toward a sociocultural model of teaching and learning. *Journal of Research in Science Teaching, 29*(8), 791-820.

Perrow, C. (1967). A framework for the comparative analysis of organizations. *American Sociological Review, 32*, 194-208.